DOCUMENT RESUME

ED 152 842 TM 007 057

AUTHOR Scheetz, James P.; Forsyth, Robert A.
TITLE A Comparison of Simple Random Sampling Versus Stratification for Allocating Items to Subtests in Multiple Matrix Sampling.
PUB DATE Apr 77
NOTE 21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, New York, April 5-7, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Achievement Tests; *Complexity Level; Item Analysis; *Item Sampling; Mathematics; *Matrices; Sampling; Secondary Education; Statistical Analysis; Statistical Data; *Test Construction; Test Items; Vocabulary
IDENTIFIERS *Item Discrimination (Tests); *Multiple Matrix Sampling; Test Length

ABSTRACT

Empirical evidence is presented related to the effects of using a stratified sampling of items in multiple matrix sampling on the accuracy of estimates of the population mean. Data were obtained from a sample of 600 high school students for a 36-item mathematics test and a 40-item vocabulary test, both subtests of the Iowa Tests of Educational Development. The results indicate that a stratified sampling of items, either by item difficulty level or by item discriminating ability, does not consistently yield more accurate estimates of the population mean than does simple random sampling. (Author/CTM)

A COMPARISON OF SIMPLE RANDOM SAMPLING VERSUS STRATIFICATION FOR ALLOCATING ITEMS TO SUBTESTS IN MULTIPLE MATRIX SAMPLING

James P. Scheetz
Health Sciences Center
University of Louisville
Louisville, Kentucky 40201

and

Robert A. Forsyth
University of Iowa
Iowa City, Iowa 52242

Presented at the annual meeting of the
National Council on Measurement in Education
April 5-7, 1977
New York, N.Y.

The use of multiple matrix sampling (MMS) techniques for program evaluation purposes has been accepted by educational evaluators as a means of reducing the time and cost required to perform program evaluations (Knapp, 1972). The usual MMS procedure involves the administration of a sample of items drawn from a larger item universe to a sample of respondents drawn from a population of respondents. On the basis of this information, estimates of population parameters [usually the mean (μ) and variance (σ^2)] are obtained and used as part of the evaluation data. It is desirable, of course, to estimate these parameters as accurately as possible.

Many investigators have attempted to identify procedures for selecting a MMS procedure that will provide the most accurate estimates of the population parameters [See, for example, Shoemaker (1970a, 1970b, 1971, 1972, 1973), Knapp (1972), and Barcikowski (1972, 1974)]. The majority of these investigations have been concerned with the selection of a set of design parameters [i.e., the number of subtests (t), the number of items per subtest (k), and the number of examinees per subtest (n)] that will yield accurate estimates of the mean and variance. Relatively few investigations have examined the effects of using stratified random sampling of items rather than simple random sampling of items on the accuracy of these estimates. Shoemaker (1973) has suggested that when using MMS, items should be stratified by difficulty level rather than content area. Myerberg (1975) stratified items by difficulty level and then used MMS procedures to draw samples from a computer generated data base. He found that stratification by

item difficulty did not consistently result in more stable estimates of the standard errors of $\hat{\mu}$ and $\hat{\sigma}^2$ than did simple random assignment. It was found that systematic decreases in the standard error terms occurred only when concurrent stratification by item difficulty and content was used.

With regard to item discriminating ability, Barcikowski (1972, 1974) concluded that discriminating ability as measured by the biserial correlation between each item and total test score does affect the variability of the estimated mean. Samples drawn from a universe with biserial correlations in the range .05-.50 resulted in more precise estimates of the mean than did samples drawn from a universe of items with biserial correlations in the .40-.70 range. For the variance, it was found that when the biserial correlations were relatively homogeneous, the MMS procedures provided more precise estimates of the variance than traditional examinee-sampling procedures. When the biserial correlations were relatively heterogeneous, traditional examinee-sampling provided estimates which were as precise as those obtained by the MMS procedures.

The primary purpose of this study was to provide additional empirical evidence related to the effects of using stratified sampling of items in MMS on the accuracy of estimates of the population mean. Two methods of stratification were examined: (1) stratification by item difficulty; and (2) stratification by item discriminating ability.

PROCEDURES

The method of analysis used in this study was the post mortem approach in which samples are drawn from a data base with known parameters. Sample estimates of the parameters of interest are then

computed and compared to the known data base values.

Description of the Data Base

The items used in this study were from two subtests of the Iowa Tests of Educational Development (ITED): mathematics (36 items) and vocabulary (40 items). These two subtests were chosen primarily because the distributions of scores on these subtests were known to be relatively different. Also, the items on each subtest are independent, (i.e., the response to a given item is not dependent on the responses to other items).

During 1971, 16,819 ninth grade students in Iowa took the math test (Form X-6). A 1 in T systematic sample was drawn to give a data base with N = 600. For this data base of 600 the following values were found: $\mu = 11.623$, $\sigma^2 = 30.385$, $\sigma = 5.512$, skewness = .965, and kurtosis = 3.680. The range was 32 with a minimum score of one and a maximum score of 33. The reliability (KR 20) for the data base was .779. The distribution was positively skewed and leptokurtic.

The second data base was derived from the scores of 13,821 eleventh grade students in Iowa who took the vocabulary test (Form X-6) in 1973. A 1 in T systematic sample was drawn to construct a data base with N = 600. For this data base the following values were found: $\mu = 22.682$, $\sigma^2 = 88.328$, $\sigma = 9.398$, skewness = -.076, and kurtosis = 1.971. The range was 37 with a minimum score of three and a maximum score of 40. The reliability (KR 20) for the data base was .923. The distribution was slightly negatively skewed and platykurtic.

Sampling Procedures

Two methods of sampling were employed to assign items to subtests:

(1) simple random sampling and (2) stratified random sampling. The data for stratifying the items according to difficulty level came from norms for the state of Iowa for the year preceeding the sample data collection. For the math test, the normative data came from the 1970 administration and for the vocabulary test the data came from the 1972 administration. The difficulty indices for the math test ranged from .14 to .67. For the vocabulary test the range was .22 to .78. When stratifying by item discriminating ability, the item discrimination indices were taken from item tryout information. The range for these indices was .25 to .73 for the math test and .36 to .79 for the vocabulary test. [The item discrimination indices are Flanagan indices. The high and low groups were defined on the basis of the total test score on the associated subtests of the ITED. See Flanagan (1939) for further explanation.]

Table 1 lists the sampling plans which were implemented for the mathematics test and Table 2 lists sampling plans for the vocabulary test. In these tables, t is the number of subtests, k specifies the number of items per subtest, and n is the number of respondents per subtest. IPSS indicates the number of items included in each subtest from each strata. NS specifies the number of strata and IPS indicates the number of items per strata. Those sampling plans with NS = 1 obviously did not involve stratification. The sampling of items within strata was done randomly. For example, when sampling from the math test, if IPSS = 6, NS = 2, and IPS = 18, this indicates that the 36 item math test was divided into two strata of 18 items each. For purposes of illustration assume that t = 3, k = 12, and n = 20. In this instance there are three subtests of 12 items each with six items

randomly selected from each strata composing a given subtest.

Certain relationships of interest exist among the parameters of the sampling plan:

$$(t)(k) = (NS)(IPS) = K$$

where K is the total number of items in the universe (a)

and $$(IPSS)(NS) = k \quad \text{(b)}$$

Equation (a) shows that the number of subtests multiplied by the number of items per subtest is equal to the number of strata multiplied by the number of items per strata which is equal to the number of items in the universe. Equation (b) shows that the number of items per subtest per strata multiplied by the number of strata is equal to the number of items per subtest.

Each sampling plan was implemented twice; first stratifying by difficulty level and then restratifying the items according to discriminating ability. No attempt was made to stratify the items concurrently by difficulty level and discriminating ability since in an applied evaluation setting concurrent stratification would most likely be carried out on the basis of either difficulty level or discriminating ability and content. Since the items comprising both data bases are relatively homogeneous with regard to content, stratification by content did not seem reasonable.

For all sampling plans, the sampling of the item universe was exhaustive and without replacement of items for the construction of each subtest. That is, an item assigned to a particular subtest was not returned to the item pool before constructing the next subtest. This procedure assured that every item appeared on one subtest and all

items appeared an equal number of times among the subtests which is in accordance with the guidelines proposed by Shoemaker (1973). The sampling of students from the population was done randomly so that each student's response was included for only one subtest. This was done because it is unlikely that in an applied evaluation setting a given student would respond to more than one subtest while some students would not respond to any subtest. Therefore, the sampling of both items and students was exhaustive since all items in the universe and all students in the population were utilized.

Indices of Accuracy

In most evaluation studies the major parameter of interest is the population mean, μ. This study was concerned with the accuracy of different methods of estimating μ. The actual estimate of the population mean was accomplished as follows. First, for each subtest, μ was estimated using the following formula [Shoemaker (1973), p. 27]:

$$\hat{\mu}_s = (K \sum_{i=1}^{n} \sum_{j=1}^{k} X_{ij})/nk \qquad (1)$$

where $\hat{\mu}_s$ = the estimated mean universe score for subtest s

K = the number of items in the universe

n = the number of examinees who respond to each subtest

k = the number of items per subtest

X_{ij} = the observed score for individual i on item j.

Then, the estimates from each subtest were pooled using the formula below [Shoemaker (1973), p. 38] to provide a single estimate of

the universe mean:

$$\hat{\mu}_p = \frac{\sum_{s=1}^{t} O_s \hat{\mu}_s}{\sum_{s=1}^{t} O_s} \qquad (2)$$

where $\hat{\mu}_p$ = the pooled estimate of the population mean

t = the number of subtests

$O_s = n_s k_s$; the number of observations per subtest.

The accuracy of these $\hat{\mu}_p$ estimates was examined using two somewhat related indices.

The first of these was labeled $SE(\hat{\mu}_p)$ and was computed as follows:

$$SE(\hat{\mu}_p) = \left[\frac{\sum_{p=1}^{NREPS} (\hat{\mu}_p - \bar{\hat{\mu}}_p)^2}{(NREPS - 1)} \right]^{\frac{1}{2}} \qquad (3)$$

where $\bar{\hat{\mu}}_p$ is the mean over replications of the $\hat{\mu}_p$ values.

The second index of accuracy was defined as follows:

$$\Psi_1 = \frac{\sum_{p=1}^{NREPS} |\hat{\mu}_p - \mu|}{NREPS} \qquad (4)$$

$SE(\hat{\mu}_p)$ indicates how closely the estimates of μ cluster around the average pooled estimate ($\bar{\hat{\mu}}_p$) and Ψ_1 indicates how closely the estimates of μ cluster around the true data base mean. Although these two indices are somewhat different, they are highly related since the $\hat{\mu}$ values

are unbiased estimators of μ.

Both the $SE(\hat{\mu}_p)$ and the Ψ_1 values were used to compare the stratified and nonstratified MMS plans. For all comparisons, the values of t, k, and n were held constant. For example, in Table 3, plan 1 (nonstratified) was compared with plan 2 (stratified), but plan 2 was not compared with plan 3 (nonstratified) because the latter two plans involve different values of t, k, and n.

RESULTS AND CONCLUSIONS

The results of the various sampling procedures are listed in Tables 3, 4, 6, and 7. In these tables the design parameters of the sampling plans are specified by the number of subtests (t), the number of items per subtest (k), the number of examinees per subtest (n), the total number of observations (O), the number of items per subtest from each strata (IPSS), the number of strata (NS), and the number of items per strata (IPS). The column labeled $\bar{\hat{\mu}}_p$ is the estimated value of μ pooled over replications, $SE(\hat{\mu}_p)$ indicates the standard error of $\hat{\mu}_p$ [Equation (3)] and Ψ_1 is defined by Equation (4).

The column headed STRAT indicates the method of stratification with NO indicating that the items were not stratified, DIFF indicating stratification by level of item difficulty and DISC showing that items were stratified by item discriminating ability. The last column shows the number of replications of the sampling plan (NREPS).

The results for stratification by difficulty level of the math test are presented in Table 3 and the results for the vocabulary test are contained in Table 4. As noted previously, stratified plans were compared with nonstratified plans holding t, k, and n constant. For

example, considering the math test stratified by difficulty level, plan 1 (Table 3) was compared with plan 2, plan 3 with plans 4 and 5, plan 6 with plans 7 through 9, plan 10 with plans 11 and 12, and plan 13 with plans 14 through 18. The results of these comparisons are summarized in Table 5 where the numbers in the table indicate which type of sampling plan yielded the smaller value for each of the two statistics used as a basis for comparison. For example, with $SE(\hat{\mu}_p)$ as the criterion, eight of the 13 comparisons show that the stratified plans produced smaller values of $SE(\hat{\mu}_p)$ than the comparable nonstratified plans.

The data in Tables 3, 4, and 5 do not provide conclusive evidence favoring stratification by difficulty level when assigning items to subtests. These results generally support Myerberg's (1975) contention that stratified random sampling of items by item difficulty does not necessarily result in more accurate estimates of the mean than simple random sampling of items.

The results of stratification by item discriminating ability for the math test are listed in Table 6 and the results for the vocabulary test are presented in Table 7. Table 8 summarizes the comparisons between the stratified and nonstratified designs. Again, neither type of sampling plan consistently resulted in more accurate estimates of μ. However, there was a slight tendency for the stratified sampling plans for the vocabulary test to produce more accurate estimates than the simple random sampling plans.

Concluding Statement

Generalizations from the results of this study must be made very cautiously. Only two item universes were studied. Furthermore, the number of replications used to estimate the accuracy of the two MMS procedures was extremely small for studies of this type. Nonetheless, these results do provide additional data related to the effects of using item stratification procedures in MMS. In general, these results indicate that stratified sampling of items either by item difficulty level or by item discriminating ability does not consistently yield more accurate estimates of μ than does simple random sampling.

References

Barcikowski, R. S. A monte carlo study of item sampling (versus traditional sampling) for norm construction. Journal of Educational Measurement, 1972, 9, 209-214.

Barcikowski, R. S. The effects of item discrimination on the standard errors of estimate associated with item-examinee sampling procedures. Educational and Psychological Measurement, 1974, 34, 231-237.

Flanagan, J. C. General considerations in the selection of test items and a short method of estimating the product-moment coefficient from data at the tails of a distribution. Journal of Educational Psychology, 1939, 30, 674-680.

Knapp, T. R. Item Sampling. Unpublished manuscript, University of Rochester, 1972.

Myerberg, N. J. The effect of item stratification in multiple matrix sampling. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1975.

Shoemaker, D. M. Allocation of items and examinees in estimating a norm distribution by item-sampling. Journal of Educational Measurement, 1970a, 7, 123-128.

Shoemaker, D. M. Item-examinee sampling procedures and associated standard errors in estimating test parameters. Journal of Educational Measurement, 1970b, 7, 255-262

Shoemaker, D. M. Further results on the standard errors of estimate associated with item-examinee sampling procedures. Journal of Educational Measurement, 1971, 8, 215-220.

Shoemaker, D. M. Standard errors of estimate in item-examinee sampling as a function of test reliability, variation in item difficulty indices and degree of skewness in the normative distribution. Educational and Psychological Measurement, 1972, 32, 705-714.

Shoemaker, D. M. Principles and Procedures of Multiple Matrix Sampling. Cambridge: Ballinger Publishing Company, 1973.

TABLE 1

Stratified Sampling Plans Implemented

for the Mathematics Test

Plan	t	k	n	IPSS	NS*	IPS	NREPS**
1	12	3	20	3	1	36	10
2	12	3	20	1	3	12	5
3	9	4	20	4	1	36	10
4	9	4	20	1	4	9	5
5	9	4	20	2	2	18	5
6	6	6	20	6	1	36	10
7	6	6	20	1	6	6	5
8	6	6	20	2	3	12	5
9	6	6	20	3	2	18	5
10	4	9	20	9	1	36	10
11	4	9	20	1	9	4	5
12	4	9	20	3	3	12	5
13	3	12	20	12	1	36	10
14	3	12	20	1	12	3	5
15	3	12	20	2	6	6	5
16	3	12	20	3	4	9	5
17	3	12	20	4	3	12	5
18	3	12	20	6	2	18	5

* Sampling plans with NS=1 did not involve stratification.

** The difference in NREPS between those plans with NS=1 and NS>1 is due to the need to reduce computer costs.

TABLE 2

Stratified Sampling Plans Implemented

for the Vocabulary Test

Plan	t	k	n	IPSS	NS*	IPS	NREPS
1	4	10	20	10	1	40	5
2	4	10	20	1	10	4	5
3	4	10	20	2	5	8	5
4	4	10	20	5	2	20	5
5	5	8	20	8	1	40	5
6	5	8	20	1	8	5	5
7	5	8	20	2	4	10	5
8	5	8	20	4	2	20	5
9	8	5	20	5	1	40	5
10	8	5	20	1	5	8	5
11	10	4	20	4	1	40	5
12	10	4	20	1	4	10	5
13	10	4	20	2	2	20	5

* Sampling plans with NS=1 did not involve stratification.

TABLE 3

Results of Stratification by Difficulty Level

in Assigning Items to Subtests - Mathematics (μ=11.623)

PLAN				$\bar{\mu}_p$	SE($\hat{\mu}_p$)	σ_1	STRAT	NREPS
	t 12	k 3	n 20	0 720				
	IPSS		NS	IPS				
1	3		1	36	11.620	.648	.399	NO	10
2	1		3	12	12.000	.965	.846	DIFF	5
	t 9	k 4	n 20	0 720				
	IPSS		NS	IPS				
3	4		1	36	11.320	.618	.489	NO	10
4	1		4	9	11.950	.250	.356	DIFF	5
5	2		2	18	11.920	.568	.486	DIFF	5
	t 6	k 6	n 20	0 720					
	IPSS		NS	IPS					
6	6		1	36	11.495	.571	.465	NO	10
7	1		6	6	11.490	.426	.305	DIFF	5
8	2		3	12	11.810	.792	.705	DIFF	5
9	3		2	18	11.750	.539	.385	DIFF	5
	t 4	k 9	n 20	0 720					
	IPSS		NS	IPS					
10	9		1	36	11.595	.525	.450	NO	10
11	1		9	4	11.580	.920	.775	DIFF	5
12	3		3	12	12.200	.519	.577	DIFF	5
	t 3	k 12	n 20	0 720					
	IPSS		NS	720					
13	12		1	36	11.745	.728	.576	NO	10
14	1		12	3	11.220	.676	.594	DIFF	5
15	2		6	6	11.680	.843	.635	DIFF	5
16	3		4	9	12.060	.827	.726	DIFF	5
17	4		3	12	11.960	.482	.435	DIFF	5
18	6		2	18	11.310	.580	.515	DIFF	5

TABLE 4

Results of Stratification by Difficulty Level

in Assigning Items to Subtests - Vocabulary (μ=22.682)

PLAN				$\hat{\mu}_p$	SE($\hat{\mu}_p$)	$\hat{\sigma}_1$	STRAT	NREPS
	t = 4	k = 10	n = 20	O = 800				
	IPSS		NS	IPS				
1	10		1	40	22.080	1.403	1.236	NO	5
2	1		10	4	21.780	1.106	1.229	DIFF	5
3	2		5	8	21.980	1.667	1.496	DIFF	5
4	5		2	20	21.930	.629	.752	DIFF	5
	t = 5	k = 8	n = 20	O = 800					
	IPSS		NS	IPS					
5	8		1	40	22.980	.666	.524	NO	5
6	1		8	5	23.840	1.110	1.311	DIFF	5
7	2		4	10	23.070	.925	.814	DIFF	5
8	4		2	20	22.850	.796	.541	DIFF	5
	t = 8	k = 5	n = 20	O = 800					
	IPSS		NS	IPS					
9	4		1	40	22.900	.604	.524	NO	5
10	1		5	8	22.180	1.002	.896	DIFF	5
	t = 10	k = 4	n = 20	O = 800					
	IPSS		NS	IPS					
11	4		1	40	22.280	.762	.716	NO	5
12	1		4	10	22.280	.565	.516	DIFF	5
13	2		2	20	22.470	.487	.426	DIFF	5

TABLE 5

Number of Times Nonstratified Sampling Plans and Stratified (by Difficulty Level) Sampling Plans Had Lower Criterion Values

	MATHEMATICS		VOCABULARY	
	STRAT	NO STRAT	STRAT	NO STRAT
$SE(\hat{\mu})$	8	5	4	5
Ψ_1	6	7	4	5

TABLE 6

Results of Stratification by Discriminating Ability

in Assigning Items to Subtests - Mathematics (μ=11.623)

PLAN				$\bar{\hat{\mu}}_p$	SE($\hat{\mu}_p$)	$\bar{r}_1$	STRAT	NREPS
	t k n 0 12 3 20 720							
	IPSS	NS	IPS					
1	3	1	36	11.620	.648	.399	NO	10
2	1	3	12	12.040	.506	.506	DISC	5
	t k n 0 9 4 20 720							
	IPSS	NS	IPS					
3	4	1	36	11.320	.618	.489	NO	10
4	1	4	9	11.950	.624	.505	DISC	5
5	2	2	18	11.800	.341	.286	DISC	5
	t k n 0 6 6 20 720							
	IPSS	NS	IPS					
6	6	1	36	11.495	.571	.465	NO	10
7	1	6	6	11.380	.712	.605	DISC	5
8	2	3	12	11.270	.524	.461	DISC	5
9	3	2	18	11.940	.551	.526	DISC	5
	t k n 0 4 9 20 720							
	IPSS	NS	IPS					
10	9	1	36	11.595	.525	.450	NO	10
11	1	9	4	11.780	.682	.566	DISC	5
12	3	3	12	11.380	.982	.865	DISC	5
	t k n 0 3 12 20 720							
	IPSS	NS	IPS					
13	12	1	36	11.745	.728	.576	NO	10
14	1	12	3	11.130	.489	.493	DISC	5
15	2	6	6	10.760	1.080	1.034	DISC	5
16	3	4	9	12.180	.629	.727	DISC	5
17	4	3	12	11.580	.670	.456	DISC	5
18	6	2	18	11.250	.627	.535	DISC	5

TABLE 7

Results of Stratification by Discriminating Ability in Assigning Items to Subtests - Vocabulary (μ=22.602)

PLAN				$\bar{\mu}_p$	SE($\hat{\mu}_p$)	σ_1	STRAT	NREPS
	t 4	k 10	n 20	0 800				
	IPSS		NS	IPS				
1	10		1	40	22.080	1.403	1.236	NO	5
2	1		10	4	22.580	1.406	1.204	DISC	5
3	2		5	8	21.920	1.201	1.149	DISC	5
4	5		2	20	22.610	.214	.159	DISC	5
	t 5	k 8	n 20	0 800					
	IPSS		NS	IPS					
5	8		1	40	22.980	.666	.524	NO	5
6	1		8	5	23.460	.319	.778	DISC	5
7	2		4	10	22.880	.629	.511	DISC	5
8	4		2	20	23.200	.970	.664	DISC	5
	t 8	k 5	n 20	0 800					
	IPSS		NS	IPS					
9	5		1	40	22.900	.604	.524	NO	5
10	1		[redacted]	8	22.410	.708	.586	DISC	5
	t 10	k 4	n 20	0 800					
	IPSS		NS	IPS					
11	4		1	40	22.280	.762	.716	NO	5
12	1		4	10	22.490	.558	.374	DISC	5
13	2		2	20	22.070	.622	.612	DISC	5

TABLE 8

Number of Times Nonstratified Sampling Plans and Stratified (by Discriminating Ability) Sampling Plans Had Lower Criterion Values

	MATHEMATICS		VOCABULARY	
	STRAT	NO STRAT	STRAT	NO STRAT
$SE(\hat{\mu})$	8	5	6	3
Ψ_1	6	7	6	3